World Journal of Social Science Research
ISSN 2375-9747 (Print) ISSN 2332-5534 (Online)
Vol. 8, No. 1, 2021
www.scholink.org/ojs/index.php/wjssr

Original Paper

Game Theory and *the Three-Body Problem*

Helena G. Su[1*]

[1] Lynbrook High School, San Jose, US

[*] Helena G. Su, Lynbrook High School, San Jose, US

Received: December 20 Accepted: December 31, 2020 Online Published: January 12, 2021

doi:10.22158/wjssr.v8n1p17 URL: http://dx.doi.org/10.22158/wjssr.v8n1p17

Abstract

Game theory is used to analyze a variety of social situations with competing players and interests, such as competition between businesses and elections. However, the concept of game theory is rarely employed to examine literature work. This paper applies game theory to analyze decisions made in a renowned science fiction series, the Three-Body Problem. How is deterrence established between two entities that want to conquer each other? What about changing the context to many entities? In the context of a science fiction, this paper offers a new perspective that allows readers to understand the rationale behind decisions and consequent outcomes in international relations and deterrence. Additionally, this research employs a simple computer program to help test equilibria in various conditions.

Keywords

game theory, the three-body problem, science fiction, bayesian game

1.1 Introduction

The novel *The Three-Body Problem* of *Remembrance of Earth's Past* trilogy by Cixin Liu is the first Asian novel to win a Hugo Awards. In the field of Chinese science fiction, Cixin Liu has a high reputation, and his work takes Chinese sci-fi to a new height. The novel *The Three-Body Problem* describes the first contact between the alien civilization, the Trisolarans, and the Earth civilization. The sequels to *The Three-Body Problem*, *The Dark Forest* and *Death's End* continue exploring the fate of the two civilizations and expand the scope to the whole universe, including higher-dimensions and interstellar wars. The trilogy, with its broad imagination and many implicit connections with game theory, inspired me to write a research paper and build game-theoretic models around it. Its connections with game theory are clearly demonstrated in "The Dark Forest Theory" discussed in *The Dark Forest* and the "Deterrence Era" that occurs in *Death's End*. In this paper, a game modeled on the deterrence between humans and Trisolarans, and another game based on the Dark Forest Theory and their Nash Equilibria will be

explained. The central question that this paper will attempt to answer is "How does game theory shed light on the Dark Forest theory in the science fiction series *the Three-Body Problem*?"

1.2 The Dark Forest Theory

This section briefly explains some core concepts in the novel that relate to the topic of this paper. The novel *The Dark Forest* of *Remembrance of Earth's Past trilogy* describes "the Dark Forest Theory" as a metaphor for the picture of the cosmic civilization in the story:

> "The universe is a dark forest. Every civilization is an armed hunter stalking through the trees like a ghost, gently pushing aside branches that block the path and trying to tread without sound. Even breathing is done with care. The hunter has to be careful, because everywhere in the forest are stealthy hunters like him. If he finds other life—another hunter, an angel or a demon, a delicate infant or a tottering old man, a fairy or a demigod—there's only one thing he can do: open fire and eliminate them. In this forest, hell is other people. An eternal threat that any life that exposes its own existence will be swiftly wiped out".

This theory is built upon two important ideas in the book: "the chain of suspicion" and "the technological explosion". The chain of suspicion forms due to the physical distance and cultural differences between cosmic civilizations, which create distrust between civilizations. This distrust is described as follows, "If you think I'm benevolent, that's not a reason to feel safe, because...a benevolent civilization can't predict that any other civilizations is benevolent". The technological explosion is a sudden technological advancement in a short period of time. It prevents a strong civilization from being willing to befriend a weak or harmless one, since a technological explosion can make the harmless civilization more powerful and threatening in a short time. Essentially, "the chain of suspicion" and "the technological explosion" erase any possibility of building a benevolent relationship between any two civilizations. The Dark Forest Theory is extremely vital as civilizations accept it to be the truth of cosmic society. Thus, civilizations strategize their behaviors according to the theory and attack every civilization they discover. The deterrence between Trisolaran civilization and Earth civilization is built upon the Dark Forest Theory, as explained in the following section.

2.1 The Deterrence Game

The following set of three games attempts to model the situation in the "Deterrence Era" in *The Death End* of *Remembrance of Earth's Past* trilogy. In the Deterrence Era, human society deters the Trisolarans from taking over the earth and maintains a temporary relationship between the two civilizations. The three sequential games correspond to the three stages of the deterrence: pre-deterrence, deterrence, and post-deterrence. Two transitions among three games demonstrate the formation and breakdown of the

deterrence, respectively.

2.2 Pre-Deterrence Stage

Trisolaran civilization discovered the Earth and can choose to either attack or communicate peacefully with Earth civilization. The Trisolarans aren't opposed to a friendly relationship with another civilization; however, their priority is to find and take over a planet to live on because of the harsh environment on their home planet. This scenario (Figure 1) is modeled as follows:

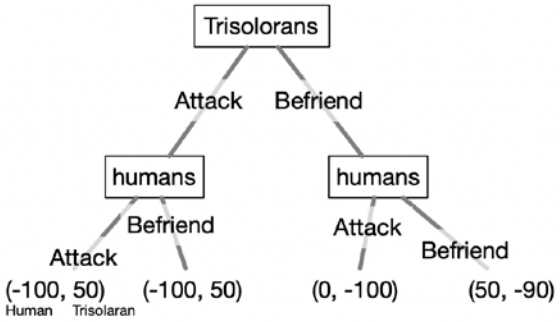

Figure 1. Pre-Deterrence

Given that Earth civilization is weaker, and that Trisolarans could easily take over the Earth, the Trisolarans would obtain a high payoff if they choose to attack, no matter whether the humans resist or not. If the Trisolarans choose to befriend humans, they will obtain a very low payoff since they would continue living on a planet with a harsh environment. If they choose to befriend humans, the Trisolarans would get a slightly higher payoff if humans also choose to befriend, compared to if humans attack (The Trisolarans' payoff of both befriending is -90, whereas the payoff of only the Trisolarans befriending is -100). However, that slightly higher payoff is still very low compared to that of taking over the Earth. The payoff for humans is more straightforward; if the Trisolarans attack, humans would be eliminated, thus a low payoff of -100. Humans would get a high payoff if they can form a friendly relationship with the Trisolarans because they can benefit from exchanges of advanced Trisolaran technology. If Trisolarans befriend and humans attack, humans would not benefit from the attack since human technology is not capable of conquering Trisolaran civilization. Note here that all payoffs in the three deterrence games are ordinal. Getting a payoff of 10 is not twice as good as that of 5; it is simply a more favorable choice.

Comparing the payoffs, it is clear that the Trisolarans will choose to attack the Earth because that is their dominant strategy. This corresponds to the plot in *The Dark Forest*; during the pre-deterrence stage, the Trisolarans decide to attack and dispatch military starships to the Earth.

2.3 Deterrence Stage

Right before the Trisolaran starships arrive on Earth, humans acquire new information and discover a new option—sending a signal that will broadcast the coordinates of the planet Trisolaris. This action is based on the idea of "the Dark Forest Theory", which assumes that every civilization will attempt to eliminate every other civilization it discovers. Thus, sending out a signal acts like a deterrence: if you attack me, I am going to send out a signal, exposing your coordinates, and others will destroy you. However, sending out the signal will also expose the coordinates of Earth, and humans are not willing to do that. The payoffs (Figure 2) are the same as the game of pre-deterrence stage, except for the "send signal" action. Since sending out the signal will eventually result in the elimination of both civilizations, the payoff is -100 for both players.

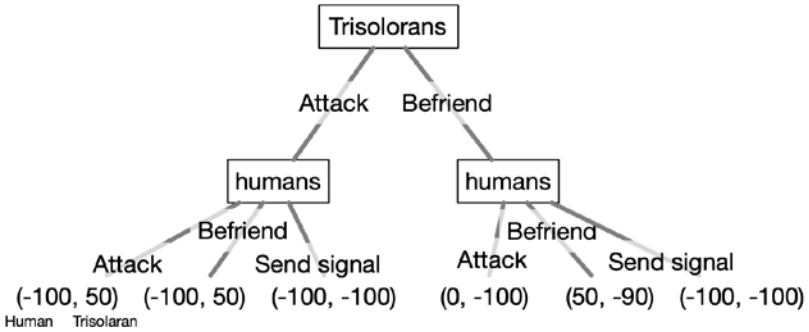

Figure 2. Successful Deterrence

When humans discover the action of broadcasting the coordinates and its significance, humans reveal this knowledge to the Trisolarans and deter them from attacking the Earth. Since the Trisolarans now know that humans are capable of broadcasting the signal and will do so if necessary, they will not choose to attack Earth. Instead, they will choose to befriend Earth civilization. It is not the best outcome, but it is better than being destroyed by other advanced civilizations in the universe. In the book, the protagonist Luo Ji threatens the Trisolarans that he will send out a signal if the Trisolarans continue to attack rather than befriend Earth civilization. Trisolarans then command their starships to return to their home planet. This scenario is similar to nuclear deterrence because both have a characteristic called the "mutual assured destruction", in which both players will be destroyed if either one of them sends out the signal or launches nuclear weapons.

2.4 Post-Deterrence Stage

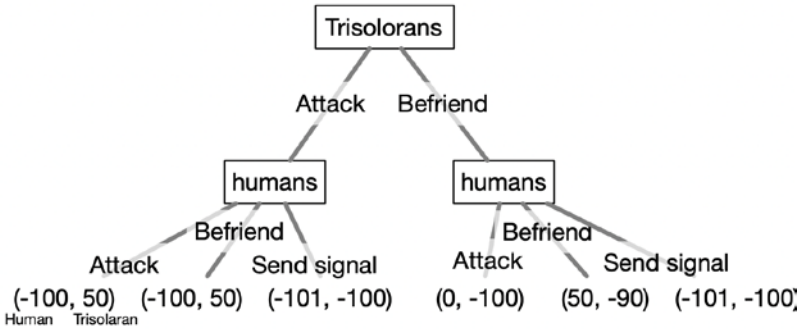

Figure 3. A Benevolent "Swordholder"

The beneficial relationship (at least beneficial for humans) doesn't last long; the Deterrence Era only lasts for several decades. The transition from Deterrence Stage to Post-Deterrence Stage (Figure 3) relates to the position of "Swordholder", who is the sole person responsible for sending out the coordinates if the Trisolarans attack. In the second game which models the deterrence stage, Luo Ji, the protagonist of *The Dark Forest*, is the Swordholder. In both games (Figure 2 and Figure 3), the Swordholder's payoff represents the payoffs of humans since the Swordholder is the only person responsible for and capable of sending out the signal. Decades later, when humans no longer consider Luo Ji as the proper person to hold such devastating power, they elect a new "Swordholder", Cheng Xin. Notably, Cheng Xin is a benevolent person, and she will not broadcast and expose the coordinates of Trisolaris even in the event of Trisolaran aggression. Thus, while other payoffs of the game stay the same, given that the Trisolarans attack, Cheng Xin's payoff of sending a signal (-101) is lower than not sending the signal.

In the book, the Trisolarans perform a psychological analysis on Cheng Xin prior to her succession to the position Swordholder, and they discover that Cheng Xin is benevolent and would not send out the signal in any situation. As soon as Cheng Xin succeeds as the second Swordholder, the Trisolarans change their action and immediately launch an attack, successfully taking over Earth.

The deterrence collapses and Earth civilization is severely hurt because humans possess less information about the actual payoffs of the Trisolarans than the information the Trisolarans know about humans. The Trisolarans deliberately hide the low payoffs of befriending humans and create the illusion that they are satisfied with the peace. In reality, the Trisolarans suffer from the harsh environment of their planet and are never satisfied with the friendly relationship because they need to take over Earth to survive. Additionally, the Trisolarans attempt to murder Luo Ji several times in order to block humans from acquiring more information about the Dark Forest Theory and thus, the option of broadcasting the coordinates of the Trisolaris to deter the Trisolarans from attacking. On the other side, humans overly

enjoy the peace between the two civilizations and gradually lose the ability to recognize the Trisolarans as menacing enemies. Humans' lack of vigilance allows them to elect a benevolent Swordholder who poses absolutely no threat to Trisolaran civilization.

3.1 The Dark Forest Theory: A Game-Theoretic Model

The game is a model of interstellar civilization interaction, and it simulates a scenario similar to "the Dark Forest Theory" in the science fiction trilogy *Remembrance of Earth's Past*. Its overall structure is similar to that of a Bayesian game in game theory. To reflect the situation in the book as well as possible, and for the sake of simplicity, the game is modeled as follows.

3.2 The Story Behind the Game

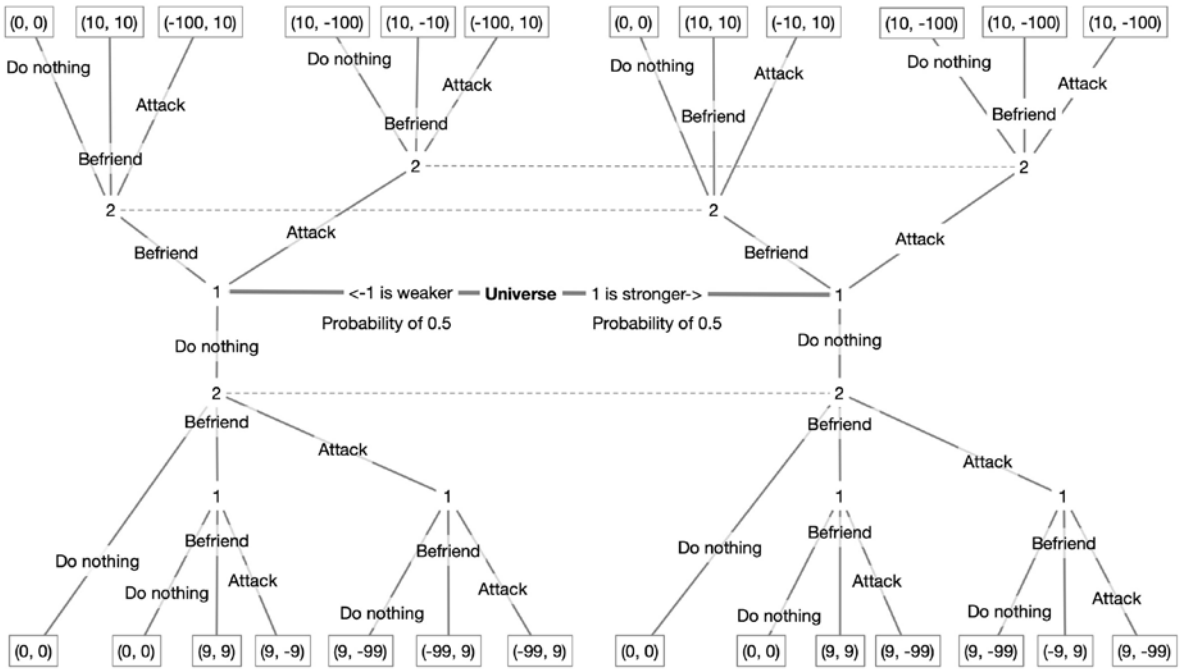

Figure 4. The Dark Forest Theory: A Game Theoretic Model

Although the game tree at first seems complicated, it will be much more understandable when the story behind this game is explained. The story goes as follows: civilization 1 (player 1) one day discovers another civilization (player 2) out there in the universe. The game now officially starts; through its observation of civilization 2, civilization 1 knows whether it is stronger or weaker than civilization 2. There are two states of the world: one in which civilization 1 is strong and civilization 2 is weak, and one in which civilization 1 is weak and civilization 2 is strong. Civilization 1 knows which state of the world

it is in, and thus knows it is in either the right branch, where civilization 1 is strong or the left branch, where civilization 1 is weak (Figure 5).

<-1 is weaker — **Universe** — 1 is stronger->

Figure 5. Two States of the World

Now, civilization 1 needs to decide what action to take with the information it has. It knows the state of the world and one universal belief that all civilizations acknowledge—survival is the most important thing. The actions civilization 1 can take, represented by the three branches (Figure 6), are to attack, befriend, or do nothing. Some possible motivations behind each action are explained in the following sentences. For example, if civilization 1 attacks civilization 2, it may successfully conquer civilization 2 and gain resources and benefits. If civilization 1 befriends civilization 2, the two civilizations might be able to form a mutually beneficial relationship and assist each other.

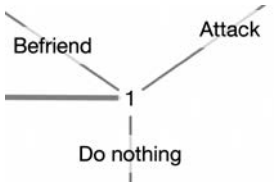

Figure 6. Three Actions of the Player 1

At this point, civilization 1 is not sure what it should do. It doesn't know anything else about civilization 2, including its beliefs, which determine what civilization 2 will think about civilization 1's strength and actions. civilization 1 can use its knowledge of the strength of the two civilizations to help determine what action it will take. For example, if civilization 1 believes that it is stronger than civilization 2, and it chooses to attack, the game will proceed to the upper left branch (Figure 7).

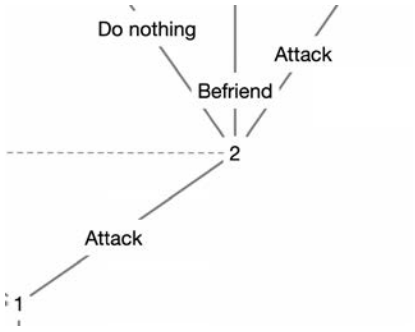

Figure 7. The Upper Left Branch

After civilization 1 makes its move, it is civilization 2's turn to act. Civilization 2 doesn't know civilization 1's strength (strong or weak). Thus, civilization 2 responds according to its belief about civilization 1's strength and observation of civilization 1's action (Figure 8).

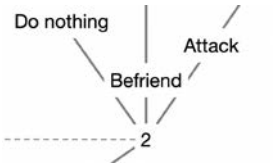

Figure 8. Actions of Player 2

Returning to civilization 1's decision and focusing on the lower branch of the game (Figure 9): other than attacking or being friendly with civilization 2, civilization 1 can choose not to expose itself to civilization 2 and simply choose to do nothing with civilization 2.

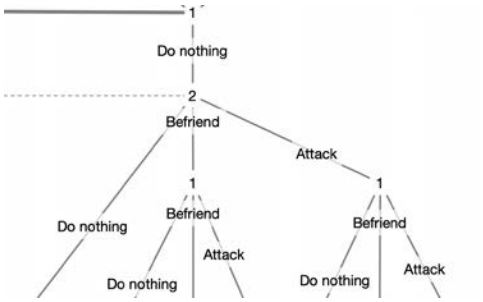

Figure 9. The Lower Branch

This is only temporary since civilization 1 knows that it is only a matter of time until civilization 2 discovers the existence of civilization 1. Because of this, the choice to do nothing essentially gives the other player the first move. If civilization 1 does nothing, civilization 2, at some point, will discover civilization 1 and repeat what civilization 1 would have gone through; it doesn't know anything about civilization 1 and has to decide what to do with civilization 1. Similarly, civilization 1 can respond to civilization 2's action of attacking or being friendly. Thus, the lower branch of the game tree is essentially a repeat of the upper branch. Unless both two civilizations are hermits and choose to do nothing to each other, there will be interactions between the two civilizations, resulting in corresponding payoffs.

3.3 The Rationale Behind the Design of the Game

The player, state of the world, actions, and beliefs of this game are briefly summarized here. There are two players in this sequential, incomplete information game. In this interstellar scenario, each player

represents a civilization, and player 1 moves first. When each of the players is making decisions, three actions are available to choose from: attack, befriend or do nothing. There are two states of the world: player 1 is stronger or player 1 is weaker. The assumption in this game is that there is a 50/50 chance that civilization 1 is strong versus weak, and both civilizations acknowledge that. Besides that, players can have additional beliefs about the other player, including what strength they have or what actions they are going to take.

The incomplete information structure attempts to model the concept of the chain of suspicion introduced in *the Dark Forest Theory*. The chain of suspicion indicates that civilizations don't know about each other's strength and strategy. In this game, player 2 doesn't know the strength of the player and neither knows the other's beliefs. In this incomplete information game, beliefs, including the universal belief that gives intuition about each other's payoff, are great tools to help players to decide what strategy they are going to play. For example, if player 2 believes that a friendly player 1 is weak, it may choose to attack because it thinks it will win. players may also take advantage of the universal belief that elimination is the worst outcome for every civilization and strategize their behaviors based on that.

The two states of the world in the game are directly determined by the strength of the two players. The first state is where player 1 is stronger than player 2, and the second state is where player 2 is stronger than player 1. The strength of player 1 is not an absolute measurement, rather, it is relative to the strength of player 2. For example, a strong player 1 means that player 1 is relatively stronger than player 2 and vice versa. Thus, in either state of the world, if a player is strong, then the other player must be weak, and vice versa. In this game, player 1 is given a first-mover advantage because it knows which player is the stronger one, or the state of the world, before it acts. That vital piece of information will help player 1 strategize its behavior and possibly determine the outcome. It is reasonable that civilization 1 is awarded a first-mover advantage because it is likely to be more technologically advanced than civilization 2, due to the fact that civilization 1 discovers civilization 2 first.

The action "do nothing" is added deliberately to represent a special strategy available to players. When player 1 discovers player 2, it can choose to attack, befriend, or simply do nothing. Both player 1's actions of attacking and befriending will immediately expose player 1 to player 2, who can respond quickly. The action of doing nothing seems purposeless, but it has one important effect: by doing nothing, player 1 keeps itself hidden from player 2 temporarily. Although player 2 will eventually discover the existence of player 1 after a period of time, player 2 cannot respond immediately if player 1 chooses to conceal itself. Note that there is a role switch: player 1 is the first mover and player 2 will respond sequentially if player 1 chooses to attack or befriend, but if player 1 chooses to do nothing, player 2 becomes the first mover and player 1 will respond sequentially. Doing nothing may be beneficial to player 1; for example, if player 1 discovers a strong player 2, player 1 might want to delay its encounter with player 2 so it chooses to hide. The case where player 1 chooses to do nothing, and then a strong

player 2 discovers and eliminates player 1, is better than the case where player 1 exposes itself to a strong player 2 and is eliminated immediately. In the former case, player 1 survives for a longer period of time since it delays its defeat or encounter with player 2. As written in the game tree, the payoff of the former case of a delayed defeat is higher than an immediate defeat. Vice versa, in the case of a friendly relationship, since players want to benefit from the friendly relationship as early as possible, the payoff of a delayed friendly relationship is lower than an immediate friendly relationship. For example, given player 2 is stronger in the AA situation (player 1 attacks then player 2 attacks) the payoff is (-100, 10), respectively. But the lower branch counterpart NAA (player 1 first does nothing, then player 2 attacks, and then player 1 attacks) yields a payoff of (-99, 9), respectively.

Some assumptions about this cosmic society shape the payoffs in the game. The most important assumption incorporated in the game is taken from the science fiction source: "Survival is the primary need of civilization" (Liu, *the Dark Forest*). Thus, no matter which civilization is, the worst outcome for them is to be eliminated; this is reflected in the payoffs in the game where -100 is assigned to the civilization that is eliminated by another civilization. The strength or state of the world directly determines the winner in the situation of both attacking. The winner, who benefits from acquiring resources from the civilization it conquers, gets a payoff of 5 if the other player chooses to counter-attack, and a higher payoff of 10 if the other player chooses to be friendly or do nothing. No matter how the second civilization responds (attack, be friendly, or do nothing) to an attack from a strong civilization, the payoff will be -100 because the second player is eliminated; the outcome, either survive or die, is all that matters. The payoffs might be extreme when compared to normal social interaction found within human societies, but they maximize simplicity while retaining the core principle of the imaginative cosmic society in *The Dark Forest*. While it was mentioned previously that the lower branch of the game tree is a repeat of the upper branch, the payoffs differ a little for the corresponding action sets. The reason that player 1 gets less payoff when it doesn't attack first is that civilizations prefer to benefit from the victory of war earlier. Vice versa, player 2 gets more payoff when it is attacked later because civilizations want to survive as long as possible. Note here that all payoffs in this game are ordinal. Getting a 10 is not twice as good as 5; it is simply a more favorable choice. Payoffs are not cardinal because it is difficult to assign each outcome accurate numerical cardinal payoffs, given that different players may evaluate each outcome differently. By giving ordinal payoffs, the preference of outcomes is more flexible and the ranking of favorability is easier to apply to all players.

3.4 Results and Equilibria

Up until now, all the essential ingredients—players, strength, and payoffs—are created in order to model the scenario of the Dark Forest Theory. The next step is to solve for the Nash equilibria. Only pure strategy Perfect Bayesian Equilibria (PBE) are explored in this game. PBE differs from normal Nash

Equilibria in that PBE requires players to have beliefs that are consistent with the equilibrium strategies of other players. The Perfect Bayesian Equilibria presented below are discovered among possible combinations of beliefs and strategies, categorized to either separating equilibria or pooling equilibria. When all four parameters are set to their default values—that is when moral equals to zero, payoffs are equal to what is written in the first game tree, three Bayesian perfect equilibria exist. Note that in all three following equilibria below, player 1's beliefs are correct about the strength of player 2 since the state of the world is included in its information set.

Table 1. The Half-half Equilibrium

State of the world	player 1 action	player 2 action	player 2 belief
player 1 is strong	attack	befriend	100% of the chance player 1 is strong
player 1 is weak	befriend	befriend	100% of the chance player 1 is weak

In the first equilibrium (Table 1), a strong player 1 will attack and a weak player 1 will befriend. Player 2 believes with a probability of 1 that player 1 is strong if player 2 observes player 1 attacking and vice versa. Player 2 will respond by being friendly no matter what action it observes.

Table 2. The Peace Equilibrium

State of the world	player 1 action	player 2 action	player 2 belief
player 1 is strong	befriend	befriend	50% of the chance player 1 is strong
player 1 is weak	befriend	befriend	50% of the chance player 1 is weak

In the second equilibrium (Table 2), both strong and weak player 1 befriends. Player 2 believes with a probability of 0.5 that player 1 is strong and another probability of 0.5 that player 1 is weak. Player 2's action in response to player 1's action of befriending is to befriend. In the case when player 1 attacks, which is a possible deviation, player 2 will respond by attacking.

Table 3. The War Equilibrium

State of the world	player 1 action	player 2 action	player 2 belief
player 1 is strong	attack	attack	100% of the chance player 1 is strong
player 1 is weak	do nothing then attack	attack	100% of the chance player 1 is weak

In the third equilibrium (Table 3), strong player 1, or the first mover, attacks and weak player 1 will do nothing. Player 2 believes with a probability of 1 that player 1 is strong if it attacks. Player 2 will respond by attacking if it observes that player 1 attacks or befriends. As stated above, the strategy of a strong player 1, or really the strategy of a strong first mover, is to attack. If player 1 does nothing, and after a period of time player 2 discovers player 1, player 2 will attack since it is strong. As mentioned earlier, in the case where player 1 does nothing, there will be a role switch of the first-mover between player 1 and player 2, and the game continues in a way that is similar to the situation where player 1 starts the game. Given player 1 does nothing, player 2 must be strong because the player 1 that does nothing is weak (remember that belief is consistent with action in these equilibria). If strong player 1 befriends, player 2 will attack. If strong player 1 does nothing, player 2 will also do nothing. Vice versa, player 1 will attack after it does nothing given player 2 befriends.

There are a few limitations in the three equilibria of the above scenarios. First, all the parameters, including moral and payoff, are set to the default values. With fixed numerical values, it is inaccurate to conclude that the three equilibria stated above exist all the time. Second, although it is assumed in game theory that players will learn and update their beliefs or actions until their beliefs are consistent with actions, which is why all the equilibria discussed here has beliefs consistent with actions. In the real world, players can have beliefs that are inconsistent with actions while they are in the process of learning and updating their strategies. Although having beliefs consistent with actions as an assumption is common in game theory, it may be a bold assumption in the book; for example, in the second equilibrium, both players believe that each other will be kind and befriend, and they act consistently with what they believe each other will do. It is a rare case in the book that both civilizations will naturally trust each other and be kind, given that they know nothing about each other. However, it is possible that such equilibrium will happen, and it is a valid and logical equilibrium where both players play their best strategies given each others' strategies. Additionally, it is much easier to identify equilibria in games with consistent beliefs and meaningful analysis can still be drawn upon in those cases.

3.5 Adding Parameters

The first limitation can be resolved by introducing parameters and assigning variables. Three parameters, moral, "victory" and "peace" are added. They are denoted as m, v, and p, respectively. As shown in the

original game tree, the parameters are in their default values, 0, 10, and 10, respectively.

Besides strength, moral, divided into benevolent and malicious categories, also plays an extremely vital role in the interaction between civilizations. Morals serve as a fundamental idea in "the chain of suspicion", which is described in the book: "You don't know whether I think you're benevolent or malicious. Next, even if you know that I think you're benevolent, and I also know that you think I'm benevolent, I don't know what you think about what I think about what you're thinking about me…This is just the third level, but the logic goes on indefinitely" (Liu, *the Dark Forest*). Essentially, even for civilizations that want to build a mutually beneficial relationship, the distrust created by incomplete information about each other's morality makes civilizations feel insecure about reaching out and befriending each other. In the book, benevolence is defined as "not taking the initiative to attack and eradicate other civilizations" and malice is defined "as the opposite". The design of moral in the game has similar ideas from the book. Note that both player 1 and player 2 have morals and they don't know about each others' morals. Malicious players will have a smaller payoff when befriending and a greater payoff when attacking, and vice versa for benevolent players. In order to better integrate morals with payoffs and better analyze the effect of morals on the outcome of the game, a numerical scale from -10 to 10 is created to determine the scale of moral of players. The reason why the range is chosen to be from -10 to 10 is that it is the smallest value that can completely offset the benefits or harm of either befriending or attacking. A non-negative moral value indicates a benevolent player who obtains more payoff when befriending and a non-positive moral value indicates a malicious player who obtains more payoff when attacking. Essentially, the bigger the value, the more benevolent the player is. A value of 0 indicates that the player is neither benevolent nor malicious, meaning that moral doesn't have an effect on the payoff. The effect of moral values on payoffs, indicated by the letter n, is presented in the game tree below. Note that moral doesn't change the final outcome for sets of action or strategies. It independently changes the numerical payoff based on the moral of each player.

Except the parameter moral, there are two parameters that correspond to two outcomes in the game. They are "victory" and "peace". The parameter "victory" is the payoff for the winning player in the situation where one or more players attack. The parameter "peace" is the payoff for a player in the situation where both players befriend. Fixed numerical payoffs in the original game tree are replaced with parameters because payoffs as parameters better describe the variability of the players' preference for each outcome, whereas fixed numerical payoffs generalize the preference of players. For example, if player 1 is exceptionally stronger than player 2, player 1 will gain limited or little benefit from the peace between itself and player 2. In this case, player 1's payoff for the peace outcome will be much lower than player 2's payoff. This scenario can be modeled more accurately if different payoffs are assigned to different players for the outcome of peace. Similar reasoning applies to both two outcomes: victory and peace. The following game tree is updated with payoffs represented as variables, denoted by the first letter of the

name of the parameter (moral as m, victory as v and peace as p). Note that the preference for delayed outcomes still exists, denoted by the -1 or +1 at the end of the payoff.

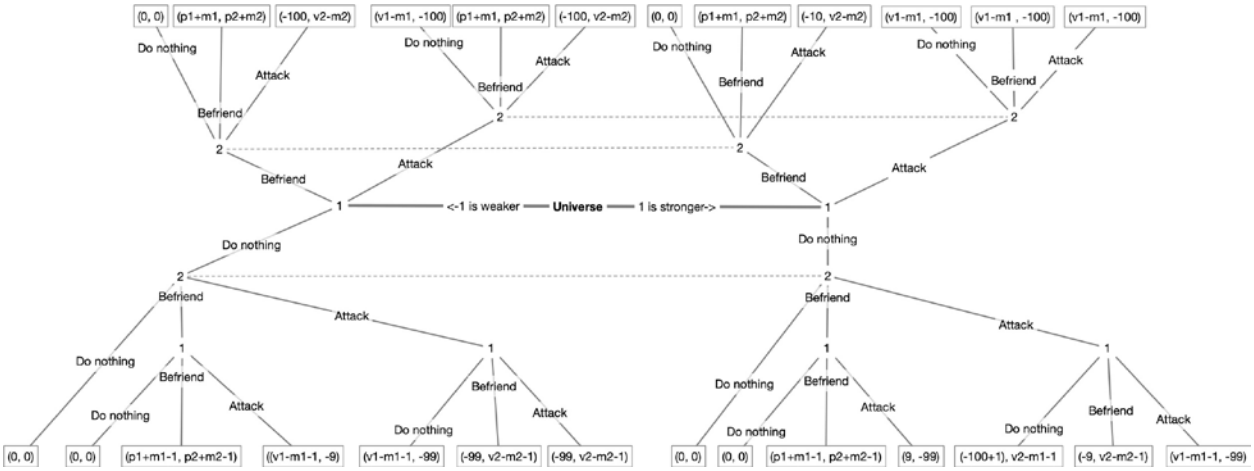

Figure 10. The Dark Forest Theory: A Game Theoretic Model with Parameters

In order to detect the effect of different values of parameters on the equilibrium, a short Python program is written. It plugs in different numerical values and quickly tests whether the equilibrium exists. It works by checking many condition statements and comparing the payoff for certain outcomes. The fundamental logic is to check each other's strategy, whether or not there is a profitable deviation. Another important use of a python program is that it can randomize the values of parameters and run the loop thousands of times to check for each given parameter whether the equilibrium exists. Randomization is suitable to simulate unknown values in a cosmic society, because the fact that there are numerous civilizations indicates that few specific statements can be used to generalize them.

3.6 Equilibria under the Effects of Parameters

In the following paragraphs, the effect of the parameters on the equilibria is discussed. Specifically, the conditions and exact numerical range of each parameter that satisfy each equilibrium are analyzed. In each of the following three paragraphs, the moral parameter and payoff parameter are discussed individually, meaning that while any one of the parameters is discussed, the other parameters are set to their default values.

Some restrictions on the parameters must be met in order to meet equilibrium 1 where the strong player 1 attacks and weak player 1 befriends, and player 2 always befriends. Given that other parameters are in their default value, in terms of moral, player 2 must be benevolent in order for its "always befriend" strategy to be the optimal strategy because player 2 doesn't have a profitable deviation if it attacks a weak player 1. Being benevolent means that player 2 has a payoff for the peace outcome that is greater than or

equal to the payoff for the victory outcome. Player 1's moral is stricter and it must be 0 to meet the equilibrium. Since player 1 knows that player 2 will befriend no matter what, the strong player 1 will attack and not befriend only when the payoff for the winning outcome is greater or equal to the payoff for the peaceful outcome. Vice versa, the weak player 2 will befriend only when the payoff for the peaceful outcome is greater than or equal to the payoff of the winning outcome. In order to satisfy both conditions, player 1's moral must be 0. Similarly, given that the moral is 0 for both players, the peace parameter for player 1 must be equal to the victory parameter for player 1, and the peace parameter for player 2 must be greater than or equal to the victory parameter for player 2.

In equilibrium 2, the peace equilibrium where both players befriend, the precondition must be that the payoff of peace must be greater or equal to victory for both players. Thus, if the payoff parameters stay in their default values, the morals for both players must be greater or equal to 0, indicating that both civilizations are not malicious. If both players' morals are set to the default value of 0, then the value of the parameter victory must be greater or equal to the value of the parameter peace for both players. If for either one of the players, the condition above isn't met, the equilibrium—the mutual befriending outcome—will no longer be valid. Connecting back to the real world, it is reasonable that a mutually befriending relationship can only exist between kind civilizations. Values of moral and payoff parameters can be freely changed, and the equilibrium will exist as long as peace is preferred over victory for both civilizations.

Equilibrium 3, the war equilibrium in which players attack each other, is somewhat the opposite of the peaceful equilibrium (equilibrium 2). Given other parameters at the default value, the moral for both players must be smaller than 0, meaning that both civilizations are malicious. This is reasonable: only when both civilizations are malicious will there be a full-scale war in which both civilizations attack. In order to meet the equilibrium, given that moral is the default value, the range of the payoff parameter is more complex. The restriction for the payoff parameter is surprisingly low, since only the parameter victory of player 2 must be greater than 0 in order to meet the equilibrium. Compared to other equilibria in which conditions of both players' parameters must be met, only one parameter of player 2 is restricted to a range in order to meet the equilibrium. The reason for this is that, given both players will ultimately attack each other, the payoff of peace no longer matters because such an outcome will never be produced. Similarly, for player 1, the payoff of victory doesn't matter because only the payoffs of AA (player 1 attacks and player 2 attacks), FA (player 1 befriends and player 2 attacks), and NAA (player 1 does nothing, player 2 attacks, player 1 attacks) are compared for a strong player 1 when it is choosing its strategies. Since it is defined that the payoff for an early victory is always bigger than that of a late victory and that elimination is the worst outcome, player 1 will never have a profitable deviation no matter what value the parameter victory is set to. For a weak player 1, it will have no victory because player 2 will always attack. Thus, the payoff for player 1's victory can be set to any value, and the equilibrium still

exists. The restriction on player 2's victory payoff is to prevent it from doing nothing. For example, when player 1 does nothing, player 2 will also do nothing if its payoff for victory is smaller than its payoff for doing nothing. Since the payoff for both players doing nothing is 0, the payoff for player 2's victory must be greater than 0 (remember, this is a delayed victory where the payoff is subtracted by 1 because player 1 does nothing first, thus player 2's victory payoff cannot be equal to 0). The most interesting discovery is that although both players need to be malicious for this equilibrium to exist, there is little restriction on the payoff parameters including peace and victory. This means that the payoff of peace could be much higher than the payoff of winning a war, and this war equilibrium would still exist, in which case both players' optimal strategy is to attack each other rather than befriending each other.

The book *The Dark Forest* states a preferred strategy of civilizations that are exceptionally advanced: "for civilizations at a certain level of technological development, attacking may be safer and less of a hassle than probing". In this scenario, the parameters can be adjusted to model the characteristics and payoffs of civilizations that are exceptionally strong. To them, the payoff for attack, or the parameter victory, is virtually 0. The book states that the utility of attacking is higher than that of befriending. Thus, the parameter peace will be set to -1. To an exceptionally strong civilization, what it will encounter is most likely a weak civilization. For player 2, the payoffs will be the default values, and the moral for both players is an integer value randomly assigned from the range -10 to 10. The equilibrium that exists most frequently (around 28%) is the war equilibrium, meaning that the strategy of an exceptionally strong civilizations is to attack, given that their opponent's moral is a random and unknown value. The peace equilibrium also appears sometimes (around 14%), although it appears less than the war equilibrium.

What if all parameters are randomized? What will be the most robust equilibrium? The range of randomization for morality is [-10, 10], for peace is [0, 10], and for victory is [0, 10], all inclusive (other values are the default numbers). Although it is a rough estimation of an overall distribution of utility for each outcome for each civilization, it still produces intuition about the three equilibria. Out of 100000 times the loop is run, the war equilibrium exists at around 27%, the peace equilibrium exists at around 10%, and the half-half equilibrium exists at around 1%. When adjusting the range of the morality parameter, it is also interesting that malicious players have a higher chance to meet an equilibrium than benevolent players do.

4.1 Conclusion

In the paper, two scenarios—Deterrence Era and the Dark Forest Theory—are modeled and analyzed in a game-theoretic way. The Nash Equilibria explains some motives and changes of actions by the Trisolarans during the Deterrence Era. Three equilibria are found in the Dark Forest Theory game, and each equilibrium illustrates a distinctive outcome. Other than Nash Equilibria, it is important to note that game theory can be extremely fun when it is used to analyze subjects that you enjoy (Note 1).

Acknowledgement

Lastly, I would like to express my deep gratitude to my research mentor, Dr. Mark Foley, a professor of economics at Davidson College. His insights and patience guided me throughout the research.

Reference

Liu, C., & Liu, K. (2014). *The three-body problem: Rememberance of Earth's past*. New York: Tor Books.

Note

Note 1. T is exciting to hear that Netflix enlisted renowned producers to adapt *The Three-Body Problem* trilogy for television.